Ideal representations in a similarity space

Wouter Voorspoels (wouter.voorspoels@psy.kuleuven.be),

Wolf Vanpaemel (wolf.vanpaemel@psy.kuleuven.be),

Gert Storms (gert.storms@psy.kuleuven.be)

Department of Psychology, Tiensestraat 102
3000 Leuven, Belgium

Abstract

The present study provides an empirical evaluation of the ideal representation view of concept representation. We compared the ideal representation view with the more established exemplar and prototype views both in common taxonomic categories and in ad hoc categories. All three views are modeled based on underlying spatial similarity representations. Results suggest that the ideal representation is the better representation in ad hoc categories, and that the exemplar model is the better representation in the common taxonomic categories.

Keywords: concepts; category representation; computational models of concept representation; typicality; ad hoc concepts

An important and robust observation in concept representation research is that not all members of a category are equally representative of the category. For example, while a platypus is a mammal, it is not a good example of a mammal. It has many features that do not fit our image of what a mammal should be like: it has webbed feet, a beak and it lays eggs. A cow on the other hand, is a good example of a mammal to most people. In the same way, a spoon is a bad example of the category *weapons*, and a gun is a good example.

Previous research suggests that people are in agreement as to what are representative, good examples of a certain category and which examples are not (Rosch & Mervis, 1975). This graded membership structure is often referred to as the typicality gradient and has been reliably observed in a broad range of natural language categories, including common taxonomic categories (e.g. De Deyne et al., 2008) and ad hoc categories, such as goal derived categories (Barsalou, 1983, 1985)

Typicality is assumed to be closely linked to the representation of a concept (e.g., Murphy, 2002; Rosch, 1978). Theories of concept representation should therefore be able to explain the observation of a typicality gradient. The observation of a typicality gradient in different kinds of categories however, does not necessarily imply that the same processes and the same kind of concept representation underlies typicality judgments. The present study aims at evaluating different views on concept representation in different kinds of categories.

Kinds of concept representations

Two contrasting views on category representation have dominated the computational research on categories and concepts, each giving a different account of the graded internal structure of categories. In both approaches typicality is related to similarity of a category member to the category representation. The two views differ in what the category representation is assumed to consist of.

On the one hand, the prototype view states that a category is represented by an abstract summary representation, referred to as the prototype (e.g., Hampton, 1979; Posner & Keele, 1968). In this view, the concept *vehicle* is a represented by a summary of what vehicles are like on average, abstracted from specific instances of vehicles, containing information such as 'moves people or cargo from point A to point B'. The typicality of *car* for the category *vehicle* then is the similarity of *car* to this abstract prototype.

On the other hand, the exemplar view proposes that a category is represented by previously encountered instances of the category, instead of an abstract summary (e.g., Brooks, 1978; Medin & Shaffer, 1978). According to this view, typicality is conceptualized as the summed similarity of a category member to all stored members of the category. For example, the concept *vehicle* consists of memory traces of previously encountered instances of vehicles, such as *train*, *plane* and *metro* (i.e. member-categories at a lower level of abstraction). The typicality of *car* is then its summed similarity to all stored instances of *vehicle*.

Barsalou (1985) has proposed a third approach to account for the typicality gradient. Focusing on ad hoc categories – categories constructed ad hoc to serve a specific purpose, for example *things you rescue from a burning house* or *things you eat when on a die* – he proposed the idea of an ideal representation. Like a prototype representation, an ideal representation is a summary representation. Unlike a prototype which is based on average, central tendency values on the stimulus dimensions, an ideal contains extreme values on relevant dimensions. For example, a typical member of the category *things to eat when on a diet* has an extreme value on the ideal dimension 'fat percentage' – typical examples being at the extreme low end of that dimension, with a zero percentage of fat as an extreme ideal representation.

Barsalou (1985) compared a number of determinants of the typicality gradient in both common, taxonomic categories and ad hoc categories – including a prototype measure and an ideal representation measure. He found that whereas in common taxonomic categories the prototype measure was the dominant determinant of typicality, the ideal measure determined the typicality gradient of the ad hoc categories significantly.

This notion of ideal representation provides and excitingly new perspective on concept representation, but, unlike the exemplar and prototype views, it has not yet made its way into a computational model of concept representation. Recently we developed a model that attempts to translate the idea of an ideal representation to a computational model (Voorspoels, Vanpaemel & Storms, submitted) that is based on an underlying spatial similarity representation. To test whether this model is a proper translation of the notion of ideal representations, we aim at replicating the findings of Barsalou (1985) using computational models. We will compare the performance of the model that implements ideal representations to an exemplar model and a prototype model (also based on underlying similarity spaces) in common taxonomic categories and ad hoc categories. If our model is a proper implementation of ideal representations, we expect an interaction between the type of model and the kind of category. The ideal representation model should be the lesser model in the common taxonomic categories and the better model in the ad hoc categories.

Models

The models considered in the present paper are all based on underlying spatial similarity representations. In a spatial representation of a category, the members are represented by points in a M-dimensional space, and the distance between two members (i.e., between two points) is inversely related to the similarity between the two members. Such a representation is typically derived using multidimensional scaling (MDS) techniques, based on pairwise similarity data. The axes that span the similarity space of a category can be considered dimensions that are important to determine the similarity relations between members in the category. In the present study, we do not attempt to interpret the axes.

Ideal Dimension Model

The ideal dimension model (IDM) posits that an ideal dimension exists in the underlying similarity space. Each exemplar of a category has a certain value along the ideal dimension, obtained by an orthogonal projection on this dimension. The further this value is located along the dimension in the ideal direction, the more typical an exemplar is.

It is useful to think of the ideal dimension as a specific combination of (unarticulated) features. The more a member has of this combination of features, the more typical it is for the category. In the case of *things to eat when on a diet*, the ideal dimension possibly is made up by a combination of features such as fat percentage, sweetness and calories. For taxonomic categories, it is more difficult to articulate the specific combination of features that might make up the ideal. To put it somewhat trivially: a car is typical for the category of *vehicles* if it has a lot of the combination of features that make up "vehicle-ness".

Formally, the IDM assumes that judging the typicality of an item i for a category A comes down to evaluating the value of i on a certain dimension V_A. In an M-dimensional space, the typicality of item i for category A, is then given by:

$$T_{iA} = \frac{\sum_{k=1}^{M} x_{ik} x_{Ak}}{\left(\sum_{k=1}^{M} (x_{Ak})^2 \right)^{1/2}}, \qquad (1)$$

where x_{Ak} are the coordinates spanning the ideal dimension V_A, x_{ik} are the coordinates of item i, and M is the number of dimensions. We restrict x_A to be at a fixed distance from the origin. This does not pose a restriction for the ideal dimension.

The model orthogonally projects item i on the ideal dimension V_A , and returns a dimensional value relative to the origin that rises when the projection is farther in the ideal direction (i.e., the direction determined by the vector V_A). This value is considered the typicality of item i for category A.

Generalized Context Model

The generalized context model (GCM; Nosofsky, 1984, 1986) assumes that categorization decisions are based on similarity comparisons with individually stored category exemplars. Originally, the model was developed to account for categorization decisions, but it has successfully been adapted for typicality judgments (Nosofsky, 1991; Voorspoels, et al. 2008a).

Typicality of an exemplar is calculated by summing the similarity of that exemplar to all other exemplars in the category. Formally, the typicality of an exemplar i for category A is then given by:

$$T_{iA} = \sum_{j=1}^{n} \eta_{ij} , \qquad (2)$$

where η_{ij} is the similarity of exemplar i to exemplar j, with j belonging to category A.

The similarity between two exemplars is a function of the distance of the exemplars in the M-dimensional psychological space, adjusted by attentional weights – that specify which underlying dimensions are important in the similarity calculation – and a sensitivity parameter – which magnifies or shrinks the psychological space. Formally, the scaled psychological distance between two exemplars i and j is given by:

$$d_{ij} = c\left(\sum_{k=1}^{M} w_k \left| x_{ik} - x_{jk} \right|^r \right)^{1/r}, \quad (3)$$

where x_{ik} and x_{jk} are the coordinates of exemplars *i* and *j* on dimension *k*, w_k a parameter reflecting the attention weight for dimension *k*, *M* is the number of dimensions, and *c* is the sensitivity parameter. Since Euclidean distances are generally accepted to be more appropriate for integral dimensions (Shepard, 1964), we fixed *r* at 2 for the present studies.

Similarity of a stimulus *i* to another stimulus *j*, is related to psychological distance as follows:

$$\eta_{ij} = \exp\left(-d_{ij}\right), \quad (4)$$

where d_{ij} is the scaled psychological distance between exemplar *i* and *j*. The free parameters in the GCM consist of M-1 dimension weights and a scaling parameter *c*.

MDS-based Prototype Model

Within the framework of the GCM, one can easily define a prototype model (MPM; Nosofsky, 1992). Typicality of a category member then is the similarity towards the prototype of the category:

$$T_{iA} = \eta_{iP_A}, \quad (5)$$

where P_A is the prototype of category *A*. The position of the prototype in the similarity space is determined by averaging the coordinates of all category members on each axis.
The free parameters in the model are identical to the free parameters in the GCM (i.e., M-1 dimension weights and a scaling parameter).

Data

Construction of the psychological space relies on similarity data. Evaluation of the models relies on typicality data. For the common categories we used data from a recent norm study De Deyne et al. (2008). For the ad hoc categories, we collected the data. We will discuss the data for both category types in turn.

Common taxonomic categories

Eleven common taxonomic categories, from two semantic domains (animals and artifacts) were used in the present study (from de Deyne et al., 2008): *birds, fish, insects, mammals, reptiles, clothes, kitchen utensils, musical instruments, tools, vehicles* and *weapons*. The categories contain between 22 and 30 members.

Typicality measure The exemplars of each category, presented as verbal stimuli, were rated by 28 participants for goodness-of-example for the superordinate category they belonged to on a Likert-rating scale ranging from 1 for very bad examples to 20 for very good examples. The reliability of the judgments was evaluated by means of split-half correlations corrected with the Spearman-Brown formula, and ranged from .91 to .98 across the 11 categories (De Deyne et al., 2008, Table 1, p. 1033). The ratings were averaged over participants.

Similarity measure Pairwise similarity ratings were also available in de Deyne et al. (2008). Similarity of each member pair within a category was rated by 15 to 25 participants (varying across categories, not within categories). Estimated reliability of the ratings ranged from .88 and .96 across categories.

Ad hoc categories

Ten ad hoc categories were constructed, including those of Barsalou (1985): *things you put in your car, things you rescue from a burning house, things not to eat/drink when on a diet, wedding gifts, things you use to bake an apple pie, things you take to the beach, means of transport between Brussels and London, properties and actions that make you win the election, weapons used for hunting* and *tools used when gardening*.

For each of the categories, 80 participants generated at least eight members. From the resulting potential members pool, we sampled 20 to 25 members, covering the production frequency dimension.

Typicality measure The members of each category were rated for goodness-of-example by 30 participants on a Likert-rating scale ranging from 1 for very bad examples to 20 for very good examples. The reliability of the judgments was evaluated by means of split-half correlations corrected with the Spearman-Brown formula, and ranged from .94 to .98.

Similarity measure Since the members of an ad hoc category can be very divers and seemingly irrelevant to each other (e.g., tissues and candy), we did not ask participants to directly rate the similarity of each member pair within a category. Participants performed a sorting task, an often applied technique to arrive at a similarity measure for large stimuli sets (e.g., Ameel & Storms, 2006; Van der Kloot & Van Herk, 1991). We will briefly describe the procedure.

For each category, 60 participants sorted the members into piles according to whatever principle they thought was fitting, the only restriction being that there had to be more than one pile and less than the number of members in a category. Following their initial sort, they were asked to either further divide the piles they made in subgroups (when the number of piles in the initial sort was smaller than five), or to join piles together (when the number of piles was larger than five). This procedure resulted in 120 exemplar-by-exemplar matrices (on for each separate sort) for each category, each cell reflecting whether the pair was in the same pile or not. We summed the 120 matrices, arriving at one matrix per category, the summed scores in the cells reflecting the similarity between two members.

Results

The similarity measures for all 21 categories were used as input for a SAS non-metric MDS analyses, resulting in spatial representations in Dimensionalities 2 to 8. Stress values, measuring the badness-of-fit for the resulting geometric representation, showed a monotonically decreasing pattern in each category, indicating that the algorithm did not get trapped in a local minimum. Overall, the stress values dropped below .1 from Dimensionality 4 onwards for the common taxonomic categories and from Dimensionality 3 onwards for the ad hoc categories. Taking into account stress and the number of members of the categories, we will present results for the common taxonomic categories in Dimensionalities 4 to 8 and for the ad hoc categories from Dimensionality 3 to 6 (following generally used rules of thumb regarding number of dimensions and stress).

Recently, increasing attention has been drawn to the importance of a model's flexibility and complexity in model evaluation, and the necessity to penalize models that are more complex (any data pattern can be accounted for perfectly by a sufficiently complex model). Comparing the best fit a model can provide ignores this complexity, while assessing the average fit of the model across all possible parameter values balances model complexity and data fit (e.g., Pitt, Kim & Myung, 2003). This average fit is measured by the marginal likelihood. Given the differences in functional form of the GCM and IDM, the model evaluation in terms of marginal likelihoods is preferable.

The results of the model analyses are reported through model weights. The model weight of a model reflects the relative evidence that the data provide in favor of that model, within the set of all models that are evaluated. The evidence for a model is the marginal likelihood of the model – calculated by sampling the parameter space. For each sampled parameter value, one can calculate the likelihood given the prior distributions of the parameters. After a number of samples, the average of all samples will converge into an estimate of the marginal likelihood of the model.

We relied on standard uninformative priors. For the IDM, this translates to a uniform prior over all points at a certain distance of the origin. For the GCM and the prototype model, a uniform prior over the range 0 to 1 was used for the dimensional weights, adding the restriction that the dimensional weights have to sum to 1. The prior for the sensitivity parameter followed a Gamma(.001,.001) distribution.

We will first present the results of the analyses of the common categories. Then we will present the results for the ad hoc categories.

Common taxonomic categories

Figure 1 presents the model weights for all three models for the common taxonomic categories. For 9 out of 11 categories, the results are highly consistent across dimensionalities. Results are not consistent for *musical instruments* and *vehicles,* consequently making inferences regarding these categories rather difficult. We will consider the results of categories *fish* and *tools* to be consistent, since only in Dimensionality 4 they deviate from the other Dimensionalities. For *tools*, closer inspection of the underlying representation revealed that stress-values dropped below .1 from Dimensionality 5 onwards, possibly explaining the anomaly in the Dimensionality 4.

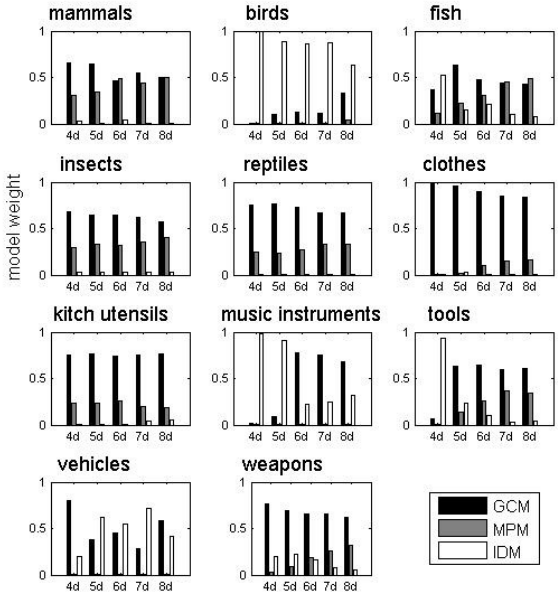

Figure 1. Model weights for the GCM, MPM and IDM for the common taxonomic categories.

It can be seen that for the 9 consistent categories, the GCM gives the better account of the typicality gradient for 8 out of 9 categories. For only 1 out of 9 categories, *birds*, the IDM clearly provides a better account. The MPM is not competitive in the present evaluation. Only for the category *fish*, it seems to provide a viable alternative in higher Dimensionalities (but even there, the MPM is not convincingly better).

In sum, the GCM seems to be the better model for the typicality gradient of the common taxonomic categories. The prototype model is never competitive, performing worse than the GCM in all categories and nearly all dimensionalities. This result confirms results of earlier comparisons between the exemplar view and the prototype view in common taxonomic concepts (e.g., Voorspoels et al. 2008) and artificial category learning (Nosofsky, 1992, Vanpaemel & Storms, 2010). The IDM possibly drives the typicality gradient of a small minority of common taxonomic categories (only *birds* in our set).

Ad hoc categories

Figure 2 presents the model weights of the three models for the ad hoc categories. For 9 out of 10 categories, the results are consistent across dimensionalities. Results are not consistent across dimensionalities for *things you take to a*

beach. Looking at the 9 consistent categories, the evidence is overwhelmingly in favor of the IDM in 7 categories. Only for the categories *hunting weapons* and *things you use when baking an apple pie* the GCM (in close competition with the MPM for the latter) is the best model. In sum, the ideal representation view indeed seems to provide a better account of the typicality gradient of ad hoc categories than the prototype and exemplar view, yet the evidence is not univocal.

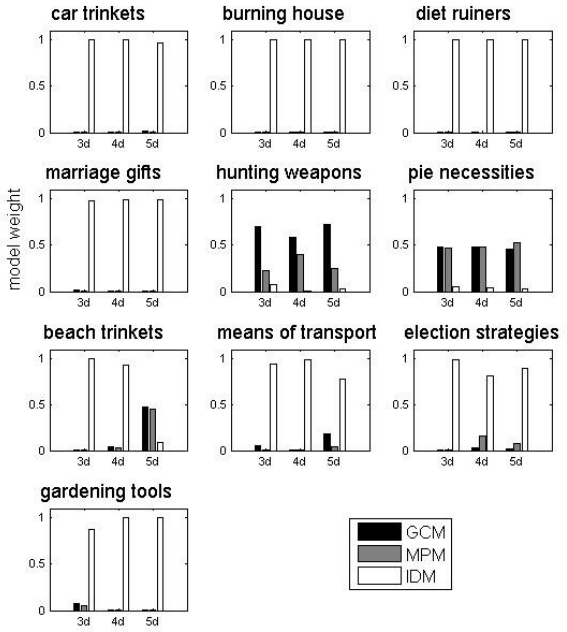

Figure 2. Model weights for the three models for the set of ad hoc categories.

The model weights reported are a relative measure of model performance, i.e., the model weight only reflects the performance of a model relative to a set of competitive models. To our knowledge however, the representational mode and the computational models used in the present study have not been applied to ad hoc categories. It is therefore informative to evaluate whether the models can give a sufficient account of the typicality gradient in absolute terms.

To this end we calculated correlations between observed and predicted typicality scores, using the optimal parameter values for each model. Results of these analyses are presented in Figure 3. It can be seen in Figure 3 that correlations rise above .6 for all categories in which the IDM is to be preferred based on the model weights, except for *properties and actions that make you win the election* and *means of transport between Brussels and London*. For the categories in which evidence based on the model weights was not in favor of the IDM, or the model weights were not consistent across dimensionalities, the optimal correlations are generally somewhat lower.

Discussion

The present study focused on the IDM, a model that provides a computational account of the notion of an ideal representation in the context of spatial similarity representations. The IDM was evaluated in its account of the typicality gradient both common taxonomic categories and ad hoc categories and compared to the GCM, arguably the most successful exemplar model, and the MPM. Following earlier findings by Barsalou (1985), we hypothesized that the IDM would have difficulty accounting for the typicality gradient of the common taxonomic categories, but that it would give a better account of the typicality gradient of ad hoc categories.

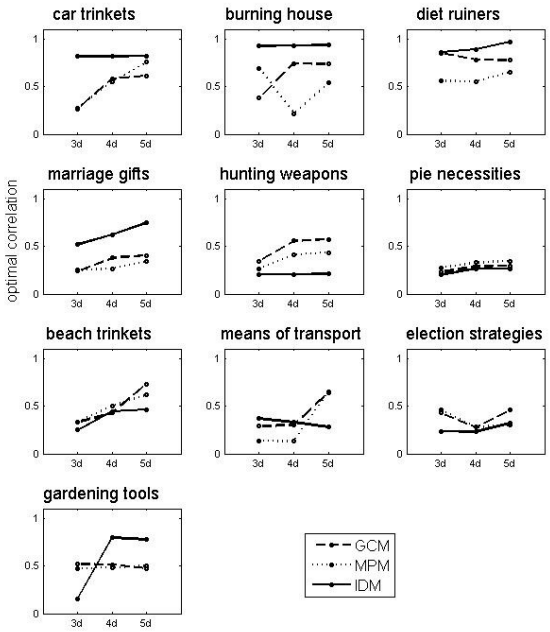

Figure 3. Optimal correlations between observed and predicted typicality ratings as a function of Dimensionality

The results supported the hypothesis. While evidence was not consistent across dimensionalities for 3 out of 21 categories, the overall pattern clearly showed the expected interaction: in the common taxonomic categories, the GCM was the better model – as can be expected based on earlier findings – and in the ad hoc categories the IDM was the better model. The evidence in any case strongly suggests that the typicality gradient of common taxonomic categories and of ad hoc categories is determined by a different representation. Moreover, the results support the reasonableness of the IDM as a formal implementation of Barsalou's (1985) notion of ideal representation.

It is unclear why this pattern broke down in 3 out of the 16 "consistent" categories. For *fish*, the IDM was the better model. In *hunting weapons* and *things you use to make an apple pie*, the GCM (MPM respectively) was the better model. Note however that for *things you use to make an*

apple pie, none of the models could give a good account of the typicality gradient in terms of optimal correlations (see Figure 3). This might suggest that the typicality gradient in this category is driven by yet another process, different from than the ones under consideration. For *hunting weapons*, the category might be considered a well-established category, rather than an ad hoc category.

To a certain extent, this study is a replication of Barsalou's work on ad hoc categories and ideal representations (Barsalou, 1985). There are, however, three crucial differences. First, we compared the ideal dimension approach to (advanced implementations of) both a prototype approach and an exemplar approach. This is important, since in this study, and in previous studies (e.g., Voorspoels et al., 2008) it is found that the exemplar approach is to be preferred over the prototype approach in concept representation.

Second, Barsalou (1985) used a priori ideals, which were generated intuitively by the researchers, for which all members of the relevant category were rated. No such instruction takes place with the IDM.

Third, Barsalou (1985) evaluated the relative contribution of different determinants of typicality, such as ideals and central tendencies, using regression analyses and a number of measures of these determinants. We tested and compared computational models of typicality that are derived from assumptions concerning concept representation. Importantly, we developed a computational model that introduces the notion of ideal representation to the context of underlying spatial representations in an intuitive way. An important finding of the present study is that the IDM indeed can be considered a computational model of ideal representations, which can be usefully applied in the further investigation of differences between concepts in terms of concept representation.

Acknowledgments

The Research in this article is part of research project G.0281.06 sponsored by the Belgian National Science Foundation – Flanders, given to the third author. We want to thank Sander Vanhaute for his help with collecting data.

References

Ameel, E., & Storms, G. (2006). From prototypes to caricatures: Geometrical models for concept typicality. *Journal of Memory and Language, 55*, 402-421.

Barsalou, L.W. (1983). Ad hoc categories. *Memory & Cognition, 11*, 211-227.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11, 4*, 629 - 654.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavioral Research Methods, 40*, 1030-1048.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, **18**, 441-461.

Medin, D. M., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.

Murphy, G. L. (2002). *The Big Book of Concepts*. Cambridge: MIT Press.

Nosofsky, R. N. (1984). Choice, Similarity, and the context model of classification. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10*, 104-114.

Nosofsky, R. N. (1986). Attention, Similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.

Nosofsky, R. N. (1991). Typicality in logically defined categories: Exemplar-similarity versus rule instantiation. *Memory & Cognition, 19*, 131-150

Nosofsky, R., N. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy, & S. M. Kosslyn (Eds.), *Essays in honor of William K. Estes*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Pitt, M., A., Kim, W., & Myung, J. (2003). Flexibility versus generalizability in model selection. *Psychonomic Bulletin & Review, 10,* 29-44.

Posner, M.I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 3*, 392-407.

Rosch, E. (1978). Principles of categorization. In E. Rosch and B. B. Lloyd (Eds.), *Cognition and categorization (pp. 27-48)*. Hillsdale, NJ: Erlbaum.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573-605.

Shepard, R. N. (1964). Attention and the metric structure of stimulus space. *Journal of Mathematical Psychology, 1*, 54-87

Van der Kloot, W. A., van Herk, H. (1991). Multidimensional scaling of sorting data: A comparison of three procedures. *Multivariate Behavioral Research, 26 (4)*, 563-581.

Vanpaemel, W., & Storms, G. (in press). Abstraction and model evaluation in category learning. Behavior Research Methods.

Voorspoels,W., Vanpaemel, W., Storms, G. (2008). Exemplars and prototypes in natural language concepts: a typicality based evaluation. *Psychonomic Bulletin & Review, 15, 3*, 630-637.

Wagenmakers, E. J., & Farrel, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11*, 192-196.